Kemal Derya

Worcester, MA kderya@wpi.edu 857-234-5283 kemalderya.github.io kemal-derya

SUMMARY

Ph.D. Candidate in Electrical and Computer Engineering (expected December 2026) with research at the intersection of security, safety, and privacy for machine learning systems. Published at IEEE SaTML, ACM ASIACCS, and IEEE EuroS&P, with experience designing adversarial robustness evaluations, safety and alignment failure analysis, guard model evasion, and cryptographic protocol security tools. Proven track record of finding vulnerabilities in deployed software (CVE-2024-5288). Strong software development background in Python and C/C++, with experience building large-scale ML evaluation systems and security-critical engineering pipelines.

PROFESSIONAL EXPERIENCE

Vernam Lab, Worcester Polytechnic Institute *August 2022 – Present*
Ph.D. Researcher *Worcester, MA*

- Design and build software systems for ML security and safety research, including adversarial attack optimization and safety evaluation frameworks in Python (PyTorch, Hugging Face) and C/C++
- Developed adversarial suffix optimization methods targeting LLM alignment and guard models, working on hidden-state representations, refusal directions, and generation-time safety behavior
- Discovered a vulnerability in wolfSSL's TLS 1.3 implementation enabling cryptographic key extraction, coordinated responsible disclosure resulting in CVE-2024-5288
- Initiated and led multiple parallel research agendas, collaborating with advisors and cross-functional co-authors; authored and co-authored 6 research papers at top security and ML venues

Analog Devices Inc. *January 2025 – August 2025*
Systems Engineering Intern *Boston, MA*

- Designed and built automated validation and compliance software for security-critical production systems using Python, CI/CD pipelines, and structured reporting
- Developed scalable verification tooling translating complex security and privacy requirements into automated testing workflows, improving reproducibility and audit readiness
- Triaged and debugged system issues across validation infrastructure, strengthening experience in code health, diagnosis, and resolution for high-assurance products

CISEC, Sabanci University *January 2020 – June 2022*
Master's Researcher *Istanbul, Turkey*

- Designed and implemented configurable hardware accelerators and software prototypes for post-quantum cryptographic schemes (NTT/FFT, modular arithmetic), developing privacy-preserving cryptographic infrastructure in Verilog and Python
- Built performance benchmarks and evaluation tools across multiple parameter settings, strengthening software development and experimental methodology skills

EDUCATION

Worcester Polytechnic Institute *August 2022 – Present*
Ph.D. Candidate *in Electrical and Computer Engineering (GPA: 4.00/4.00)* *Worcester, MA*

- Research focus: Security, safety, and privacy of ML systems; adversarial robustness; LLM safety evaluation; cryptographic protocol analysis; hardware/software vulnerability research
- Published at IEEE SaTML 2025, ACM ASIACCS 2025, and IEEE EuroS&P 2025; initiated and owned research agendas across multiple concurrent projects

Sabanci University *January 2020 – July 2022*
M.S. *in Electronics Engineering (GPA: 3.85/4.00)* *Istanbul, Turkey*

- Thesis on high-performance cryptographic hardware for post-quantum schemes; built FPGA prototypes in Verilog and Python reference implementations for privacy-preserving cryptographic operations

Sabanci University *Sept 2014 – June 2019*
B.S. *in Electronics Engineering (with Tuition Fee Scholarship, GPA: 3.67/4.00)* *Istanbul, Turkey*

SKILLS

Programming: Python, C/C++, Verilog, CUDA C++ (basic), Bash; experience with data structures, algorithms, and performance-critical implementations

Machine Learning & Safety: PyTorch, Hugging Face Transformers, LLM safety and alignment evaluation, adversarial robustness, guard model analysis, hidden-state analysis, jailbreak evaluation, reinforcement learning for vulnerability discovery, evaluation protocol design

Security & Privacy: ML security and privacy, cryptographic protocol analysis (TLS), privacy-preserving cryptographic systems, vulnerability research, exploit development, fault injection, hardware security evaluation

Software Engineering: Large-scale evaluation systems, automated testing infrastructure, CI/CD pipelines, code review, debugging and diagnosis, reproducible research pipelines, GitHub Actions

Cryptography & Infrastructure: Post-quantum cryptography, NTT/FFT implementations, modular arithmetic, FPGA development, hardware/software co-design, SoC architecture

PROJECTS

Revisiting JBShield: Breaking and Rebuilding Representation-Level Jailbreak Defenses Preprint

- Introduced JB-GCG, an adaptive variant of GCG that breaks JBShield with 46.2% average attack success rate across five configurations.
- Proposed Representation Trajectory Verification (RTV), a Mahalanobis-based outlier detector over multi-layer representation trajectories — 0.99 AUROC against JB-GCG.
- Demonstrated that multi-layer representation consistency is a more reliable foundation for jailbreak detection than single-layer concept similarity.

Super Suffixes: Bypassing Text Generation Alignment and Guard Models Simultaneously Preprint

- Developed adversarial suffix optimization methods that simultaneously bypass LLM alignment and evade safety guard models, demonstrating compound safety failures across multiple model families
- Worked with model internals — hidden-state representations, refusal directions, and generation-time safety mechanisms — to understand and exploit alignment vulnerabilities
- Built large-scale evaluation software for measuring attack success, cross-model transferability, and detection failure, applying rigorous evaluation protocol design to ML safety research

Non-Halting Queries: Unsafe and Fixed-Point Behaviors in Large Language Models IEEE SaTML 2025

- Investigated safety failure modes in LLMs by linking pathological generation to fixed-point hidden-state dynamics, identifying new classes of unsafe model behavior
- Designed scalable evaluation protocols for assessing model behavior gaps related to safety, emphasizing measurable robustness criteria and reproducible benchmarks

FAULT+PROBE: A Generic Rowhammer-based Bit Recovery Attack (wolfSSL TLS 1.3) ACM ASIACCS 2025

- Discovered vulnerabilities enabling cryptographic key extraction in a production TLS 1.3 library, resulting in CVE-2024-5288 through coordinated responsible disclosure
- Built an end-to-end software pipeline in Python and C to reproduce faults, validate exploitability, and quantify failure modes across configurations
- Developed a generic bit-recovery methodology with implications for securing privacy-preserving cryptographic implementations against hardware fault attacks

μRL: Discovering Transient Execution Vulnerabilities Using Reinforcement Learning Preprint

- Co-authored research applying reinforcement learning to automate discovery of microarchitectural vulnerabilities, demonstrating RL-driven approaches to security testing at scale
- Contributed to experiment design, reward shaping, and evaluation, gaining hands-on experience with RL

methods applied to real-world safety-critical problems

LeapFrog: The Rowhammer Instruction Skip Attack IEEE EuroS&P 2025

- Co-authored research demonstrating Rowhammer-based instruction skipping that bypasses hardware and software security countermeasures
- Contributed to fault injection experiments and exploit validation on real hardware, strengthening expertise in infrastructure-level security evaluation

CoHA-NTT: A Configurable Hardware Accelerator for NTT-based Polynomial Multiplication *Microprocessors and Microsystems*, 2022

- Designed and implemented a configurable FPGA accelerator for cryptographic operations central to post-quantum and privacy-preserving schemes, optimizing throughput and flexibility

PUBLICATIONS & PRE-PRINTS

- **K. Derya**, B. Sunar, 2026. "Revisiting JBShield: Breaking and Rebuilding Representation-Level Jailbreak Defenses." arXiv:2605.03095 .
- A. Adiletta, K. Adiletta, **K. Derya**, B. Sunar, 2025. "Super Suffixes: Bypassing Text Generation Alignment and Guard Models Simultaneously." arXiv:2512.11783.
- G. Hammouri, **K. Derya**, and B. Sunar. "Non-Halting Queries: Exploiting Fixed Points in LLMs." *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2025.
- **Kemal Derya**, M. Caner Tol, and Berk Sunar. "FAULT+PROBE: A Generic Rowhammer-based Bit Recovery Attack." *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security (ASIA CCS '25)*, 2025.
- M. C. Tol, **K. Derya**, and B. Sunar. "μRL: Discovering Transient Execution Vulnerabilities Using Reinforcement Learning." arXiv:2502.14307.
- A. Adiletta, M. C. Tol, **K. Derya**, B. Sunar, and S. Islam. "LeapFrog: The Rowhammer Instruction Skip Attack." *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2025.
- **Kemal Derya**, Ahmet Can Mert, Erdinç Öztürk, and Erkay Savaş. "CoHA-NTT: A Configurable Hardware Accelerator for NTT-based Polynomial Multiplication." *Microprocessors and Microsystems*, 2022.